Integration of accessibility data from structure probing into RNA-RNA interaction prediction

—

Supplementary material

Milad Miladi[1], Soheila Montaseri[1], Rolf Backofen[1,2], Martin Raden[1]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany, and

[2]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany

Contents

1 Integrating SHAPE data into accessibility-based RNA-RNA interaction prediction

Given two RNA molecules with nucleotide sequences $S^1, S^2 \in \{A, C, G, U\}^*$, we define interaction I between S^1 and S^2 as a set of inter-molecular base pairs (i.e. $I = \{\ (i,j)\ |\ i \in [1, |S^1|] \wedge j \in [1, |S^2|]\}$), that are complementary (i.e. $\forall (i,j) \in I : \{S^1_i, S^2_j\} \in \{\{A,U\},\{C,G\},\{G,U\}\}$) and non-crossing (i.e. $\forall (i,j) \neq (i',j') \in I : i < i' \rightarrow j > j'$). Furthermore, any position forms at most one inter-molecular base pair (i.e. $\forall (i,j),(i',j') : i = i' \leftrightarrow j = j'$). For any interaction I, the hybridization energy $E^{hyb}(I)$ can be computed using a standard Nearest-Neighbor energy model (Turner and Mathews, 2010).

The accessibility-based free energy of an interaction I is defined by

$$E(I) = E^{hyb}(I) + ED^1(I) + ED^2(I), \tag{1}$$

where the $ED^{1,2}(\geq 0)$ terms represent the energy (penalty) needed to make the respective interacting subsequences of $S^{1,2}$ unpaired/accessible (Mückstein *et al.*, 2006; Raden *et al.*, 2018; Wright *et al.*, 2018).

To compute ED terms, we need the left-/right-most base pair of I given by $(l^1, r^2) = \arg\min_{(i,j)\in I}(i)$ and $(r^1, l^2) = \arg\max_{(i,j)\in I}(i)$, respectively. Both base pairs define the interacting subsequences, i.e. $S^1_{l^1..r^1}$ and $S^2_{l^2..r^2}$. Based on that, the penalty terms are given by

$$ED^*(I) \quad = \quad -RT \log\left(Pr^{ss}(S^*_{l^*..r^*})\right) \qquad \text{with } * \in \{1,2\}, \tag{2}$$

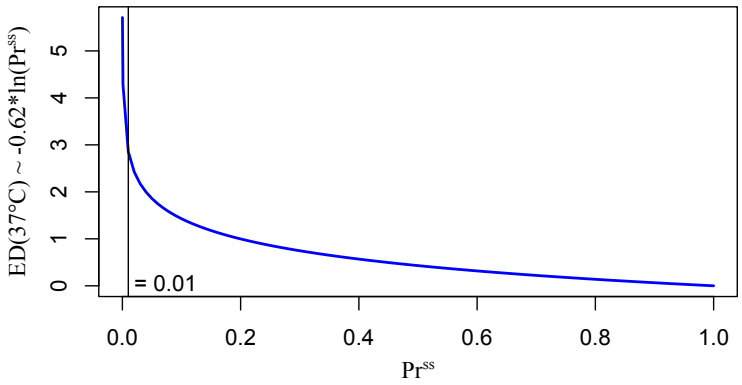
where R is the gas constant, T is the temperature, and Pr^{ss} denotes the unpaired probability of a given subsequence, which can be efficiently computed (Bernhart *et al.*, 2006).

As discussed above, SHAPE reactivity data can be incorporated into thermodynamic prediction tools via pseudo energy terms (Deigan *et al.*, 2009) as has been integrated in Vienna RNA package(Lorenz *et al.*, 2016). The latter enables SHAPE-guided computation of unpaired probabilities, i.e. the Pr^{ss} terms from Eq. 2. While SHAPE-guided energy evaluations can not be compared to unconstrained energy values (due to the pseudo-energy terms), unpaired probabilities are compatible, since they are reflecting the accessible structure space rather than individual structures. Thus, SHAPE-constrained Pr^{ss}_{SHAPE} values can be directly used within the ED computation (Eq. 2), which provides a constrained accessibility-based interaction energy (Eq. 1) without further methodical changes. This approach is implemented in the recent version of IntaRNA e.g. available via Bioconda (Grüning *et al.*, 2018).

IntaRNA interfaces all three pseudo-energy conversion methods (Zarringhalam *et al.*, 2012; Deigan *et al.*, 2009; Washietl *et al.*, 2012) currently implemented within the Vienna RNA package. We observed best performance with Zarringhalam's method (data not shown), which we attribute to the incorporation of reactivity on both paired and unpaired terms and was used for this study

Note, since unpaired probabilities are incorporated on negated log-scale only (compare Eq. 2), small or intermediate changes of high probability values are

expected to show only minor effects in the RNA-RNA interaction prediction but should still guide or fine-tune the prediction, see Fig. 1. On the contrary, if the Pr^{ss}_{SHAPE} values deviate much (orders of magnitude) from Pr^{ss} (e.g. rendering presumably unpaired regions inaccessible since they might be already blocked by other substrates or vice versa as an indirect consequence e.g. of binding) or if the probability values are very small, strong prediction effects are expected.



Supplementary Figure 1: Relation of ED penalties and unpaired probabilities Pr^{ss} at temperature $T = 37°\text{C}$, i.e. $RT \sim 0.62$.

2 Spot probabilities of RNA-RNA interaction sites

To assess the effect of SHAPE data, we define the *spot probability* Pr^{spot} of an interaction site of interest. A *spot* is defined by a pair of indices k, l for S^1, S^2, resp., and $Pr^{\text{spot}}(k, l)$ as the partition function quotient

$$Pr^{\text{spot}}(k,l) = \sum_{I' \in \mathcal{I}^*} exp(-E(I')/RT) / \sum_{I \in \mathcal{I}} exp(-E(I)/RT), \tag{3}$$

where $\mathcal{I}$ denotes the set of all possible interactions and $\mathcal{I}^* \subseteq \mathcal{I}$ the subset of interactions that cover the spot, i.e. position k, l are within the respective interacting subsequences[1] $S^1_{l^1..r^1}$ and $S^2_{l^2..r^2}$ (see above).

3 Sequence data and evaluation scripts

All sequences used within this study as well as the used evaluation scripts are available online at

[1]Note, interactions $I \in \mathcal{I}^*$ covering a spot at k, l do not necessarily contain the base pair (k, l), i.e. k, l or both can be unpaired.

https://github.com/BackofenLab/IntaRNA-benchmark-SHAPE

while the integration of the IntaRNA approach is available in version $\geq$ 2.2.0 at

https://github.com/BackofenLab/IntaRNA

This study used IntaRNA 2.2.0, ViennaRNA v2.4.7 and pseudo energy computation for SHAPE data following Zarringhalam *et al.* (2012).

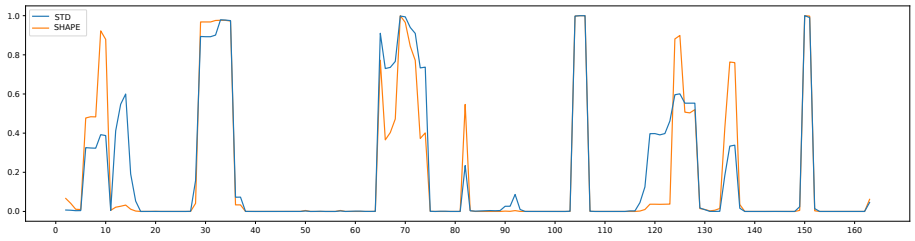
4 DMS-seq reactivity data extraction

We used the recommended settings from (Ding *et al.*, 2015) to map the sequencing reads and compute reactivities of the annotated transcripts using the Galaxy tools Afgan *et al.* (2018) Bowtie-2 and StructureFold (Langmead and Salzberg, 2012; Tang *et al.*, 2015). We selected the U1 homolog transcript bearing the largest secondary structure distance between the unconstrained and SHAPE-constrained structure prediction (using RNAfold).

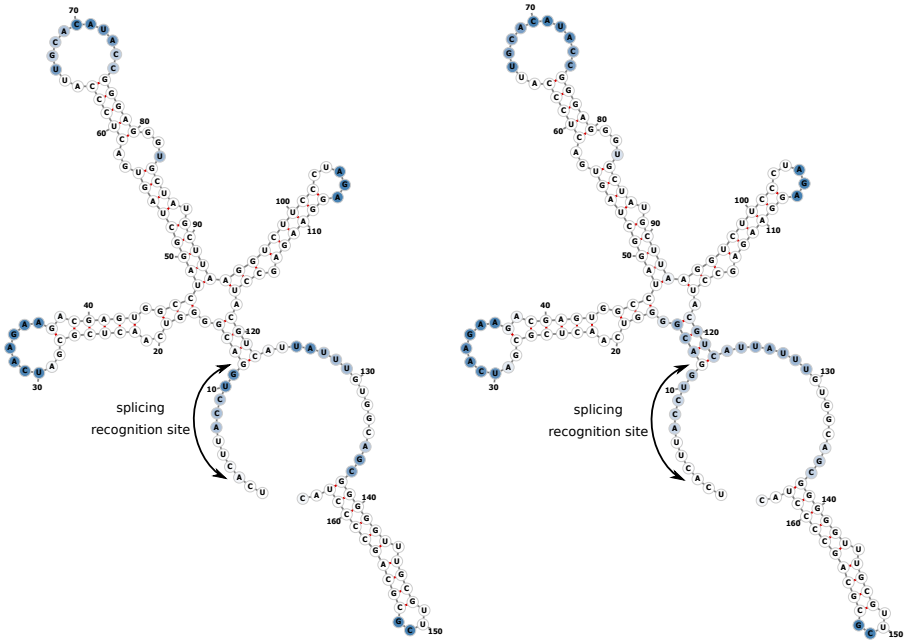
5 U1 accessibilities with and without SHAPE reactivities

Supplementary Figure 2 depicts the accessibility of the U1 RNA in terms of position-wise unpaired probabilities computed with SHAPE data and without. Without SHAPE data, the potential of the recognition site for inter-molecular interaction is underestimated, as can be seen when comparing to the SHAPE-guided unpaired probabilities. Furthermore, the subsequent region is wrongly assumed to be accessible (while known to be blocked by intra-molecular helix formation). The probabilities correspond to the 3-mer accessibilities (RNAplfold -u 3), such that for each position i the probability that 3-mer [i-2, i] is unpaired.

This can also be seen from the accessibility-annotated structure plots in Supplementary Fig. 3. The accessibility ("unpairedness") of the recognition site is much better reflected by SHAPE-guided terms compared to standard computations.



Supplementary Figure 2: Position-wise accessibility (y-axis) of U1 with SHAPE data (orange) and without (blue, STD).



(a) SHAPE (b) STD

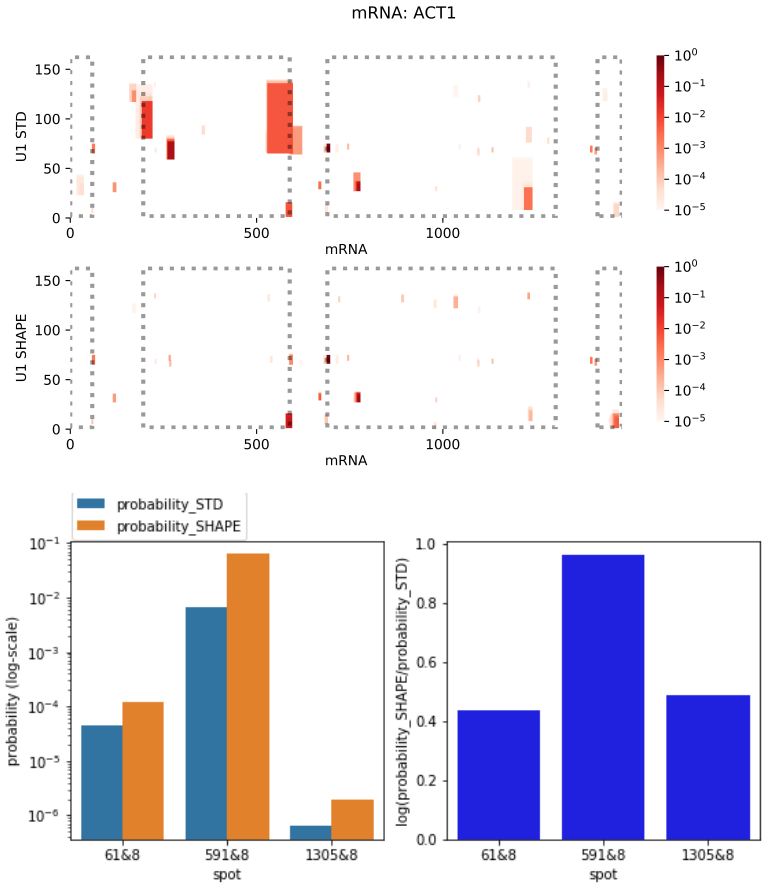
Supplementary Figure 3: Accessibilities mapped to the U1 secondary structure and color-coded, more accessible positions have darker colors. Visualization using Forna web-server (Kerpedjiev *et al.*, 2015)

6 Pre-mRNA U1 interaction predictions details

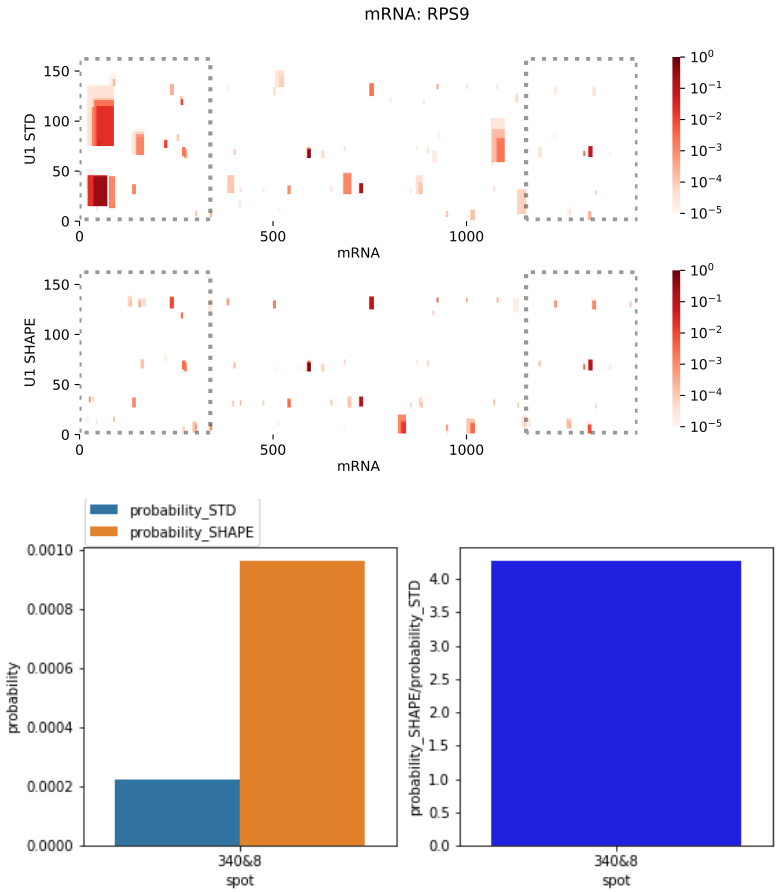
In the following, we provide the interaction prediction heatmap for all studied mRNAs. Index combinations that mark interactions of U1 with coding exons are enclosed by gray dotted boxes. The coloring represents the respective spot probabilities. Each heatmap is complemented with a visualization of the spot probabilities of U1's recognition site interacting with each CDS 5'-splice site using SHAPE data and without.

Predicted interaction sites are drawn in colored boxes where darker boxes relate to higher interaction (spot) probabilities. The x-axis corresponds to pre-mRNA indices while the y-axis represents positions of U1. For the latter, the U1 recognition site for intronic 5' splice sites is at position $\tilde{4}$-11. Thus, interactions of that site correspond to the bottom of the graph. The top graph represents interaction sites without SHAPE constraints while the bottom graph depicts the altered prediction when U1 SHAPE constraints are considered within IntaRNA's accessibility computation.

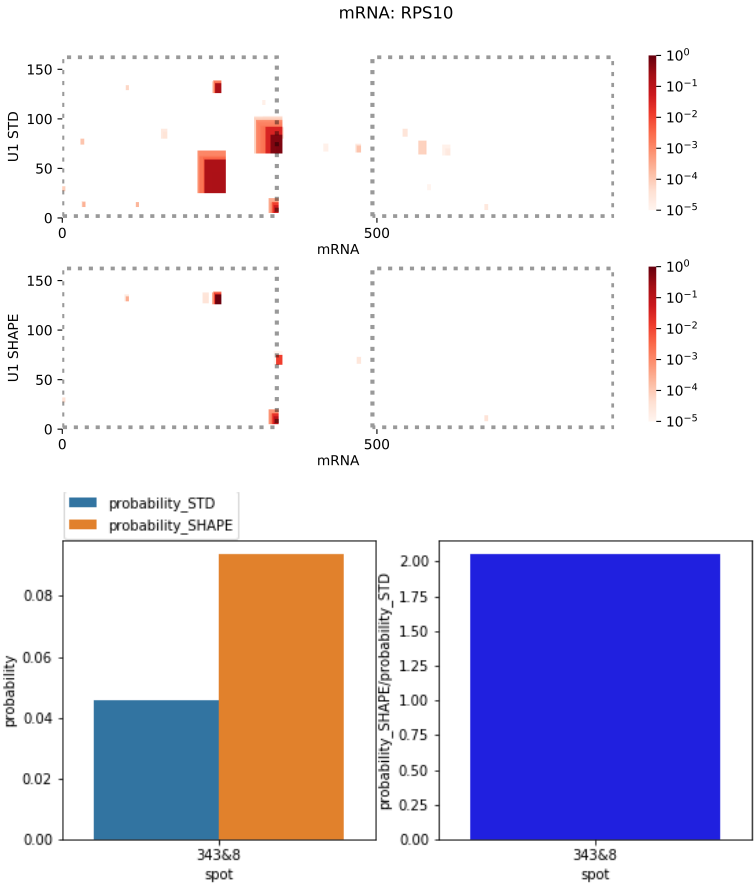
The left barplot shows the predicted interaction probabilities of all intronic 5' splice sites with the recognition site of U1 (called a spot). Blue bars represent the unconstrained probabilities while orange bars depict the probabilities when U1 SHAPE constraints are used. The right bar plot provides the ratio of both probabilities for each interaction site (spot). Please note, representations of multiple spots are in log-scaling to enable depiction.



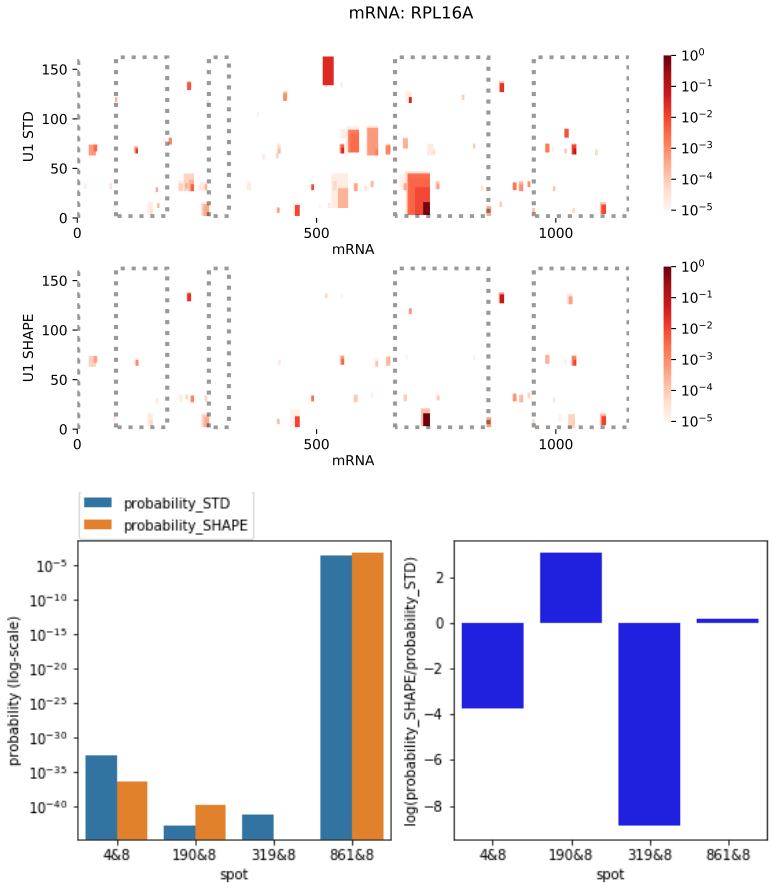
Supplementary Figure 4: U1-ACT1 interaction prediction with and without U1 probing data. Note, bar plots are in $\log_{10}$-scale.



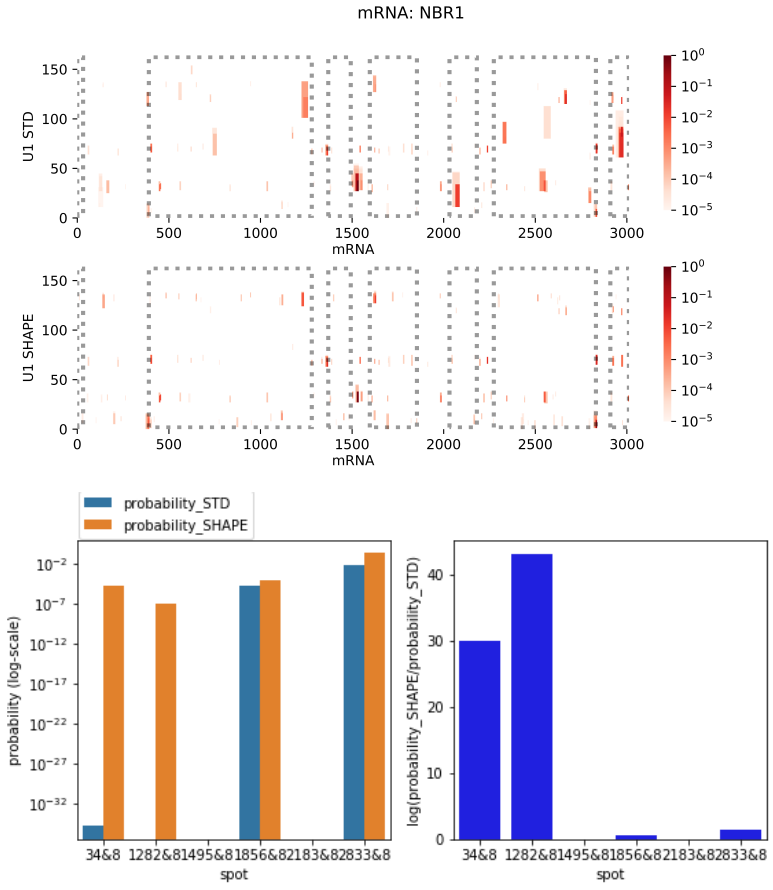
Supplementary Figure 5: U1-RPS9 interaction prediction with and without U1 probing data



Supplementary Figure 6: U1-RPS10 interaction prediction with and without U1 probing data



Supplementary Figure 7: U1-RPL16A interaction prediction with and without U1 probing data. Note, bar plots are in $\log_{10}$-scale.



Supplementary Figure 8: U1-NBR1 interaction prediction with and without U1 probing data. Note, bar plots are in $\log_{10}$-scale.

References

Afgan, E. *et al.* (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, page gky379.

Bernhart, S. H. *et al.* (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**(5), 614–615.

Deigan, K. E. *et al.* (2009). Accurate shape-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, **106**(1), 97–102.

Ding, Y. *et al.* (2015). Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nature protocols*, **10**(7), 1050.

Grüning, B. *et al.* (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*. online publication ahead of print.

Kerpedjiev, P. *et al.* (2015). Forna (force-directed rna): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**(20), 3377–3379.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4), 357.

Lorenz, R. *et al.* (2016). SHAPE directed RNA folding. *Bioinformatics*, **32**(1), 145–147.

Mückstein, U. *et al.* (2006). Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**(10), 1177.

Raden, M. *et al.* (2018). Interactive implementations of RNA structure and RNA-RNA interaction prediction approaches for example-driven teaching. *PLOS Comp Biol*, **14**(8), e1006341.

Tang, Y. *et al.* (2015). StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics*, **31**(16), 2668–2675.

Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, **38**(Database issue), D280–2.

Washietl, S. *et al.* (2012). RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic acids research*, **40**(10), 4261–4272.

Wright, P. R. *et al.* (2018). Structure and interaction prediction in prokaryotic RNA biology. *Microbiol Spectrum*, **6**(2).

Zarringhalam, K. *et al.* (2012). Integrating chemical footprinting data into RNA secondary structure prediction. *PLOS ONE*, **7**(10), 1–13.